Information-Theoretic Objective Functions for Lifelong Learning

Byoung-Tak Zhang
School of Computer Science and Engineering &
Graduate Programs in Cognitive Science and Brain Science
Seoul National University
Seoul 151-742, Korea
E-mail: btzhang@bi.snu.ac.kr

Abstract

Conventional paradigms of machine learning assume all the training data are available when learning starts. However, in lifelong learning, the examples are observed sequentially as learning unfolds, and the learner should continually explore the world and reorganize and refine the internal model or knowledge of the world. This leads to a fundamental challenge: How to balance long-term and short-term goals and how to trade-off between information gain and model complexity? These questions boil down to "what objective functions can best guide a lifelong learning agent?" Here we develop a sequential Bayesian framework for lifelong learning, build a taxonomy of lifelong-learning paradigms, and examine information-theoretic objective functions for each paradigm, with an emphasis on active learning. The objective functions can provide theoretical criteria for designing algorithms and determining effective strategies for selective sampling, representation discovery, knowledge transfer, and continual update over a lifetime of experience.

1. Introduction

Lifelong learning involves long-term interactions with the environment. In this setting, a number of learning processes should be performed continually. These include, among others, discovering representations from raw sensory data and transferring knowledge learned on previous tasks to improve learning on the current task (Eaton & desJardins, 2011). Thus, lifelong learning typically requires sequential, online, and incremental updates.

Here we focus on the aspect of never-ending exploration and continuous discovery of knowledge. In this regard, lifelong learning can be divided into passive and active learning (Cohn et al., 1990; Zhang & Veenker, 1991a; Thrun & Moeller, 1992). In passive learning the learner just observes the incoming data while in active learning the learner chooses what data to learn. Active learning can be further divided into selective and creative learning (Valiant, 1984; Zhang & Veenker, 1991b; Freund et al., 1993). Selective learning subsamples the incoming examples while creative learning generates new examples (Cohn et al., 1994, Zhang, 1994).

Lifelong learning also involves sequential revision and transfer of knowledge across samples, tasks, and domains. In terms of knowledge acquisition, the model revision typically requires restructuring of models rather than parameter tuning as in traditional machine learning or neural network algorithms. Combined with the effects of incremental and online change in both data size and model complexity, it is fundamentally important how the lifelong learner should control the model complexity and data complexity as learning unfolds over a long period or lifetime of experience.

We ask the following questions: how can a lifelong learner maximize information gain while minimizing its model complexity and costs for revision and transfer of knowledge about the world? What objective function can best guide the lifelong learning process by making trade-off between long-term and short-term goals. In this paper we focus on information-theoretic objective functions for lifelong learning, with an emphasis on active learning, and develop a taxonomy of lifelong learning paradigms based on the learning objectives.

In Section 2 we give a Bayesian framework for lifelong learning based on the perception-cycle model of cognitive systems. Section 3 describes the objective functions for lifelong learning with passive observations, such as time series prediction and target tracking. Section 4 describes the objective functions for active lifelong learning, i.e. continual learning with actions on the environment but without rewards. We also consider the measures for active constructive learning. In Section 5 we discuss the objective functions for lifelong learning with explicit rewards from the environment. Section 6 concludes by discussing the extension and further use of the framework and objective functions

2. A Framework for Lifelong Learning

Here we develop a general framework for lifelong learning that unifies active learning and constructive learning as well as passive observational learning over lifetime. We

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAR 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Information-Theoretic Objective Functions for Lifelong Learning	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Seoul National University,Graduate Programs in Cognitive Science and Brain Science,School of Computer Science and Engineering &, Seoul 151-742, Korea, ,	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Preprint, Submitted to AAAI 2013 Spring Symposium on Lifelong Machine Learning, Stanford University, March 2013,Government or Federal Purpose Rights License.

14. ABSTRACT

Conventional paradigms of machine learning assume all the training data are available when learning starts. However, in lifelong learning, the examples are observed sequentially as learning unfolds, and the learner should continually explore the world and reorganize and refine the internal model or knowledge of the world. This leads to a fundamental challenge: How to balance long-term and short-term goals and how to trade-off between information gain and model complexity? These questions boil down to ?what objective functions can best guide a lifelong learning agent?? Here we develop a sequential Bayesian framework for lifelong learning, build a taxonomy of lifelong-learning paradigms, and examine information-theoretic objective functions for each paradigm, with an emphasis on active learning. The objective functions can provide theoretical criteria for designing algorithms and determining effective strategies for selective sampling, representation discovery, knowledge transfer, and continual update over a lifetime of experience.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**9**	

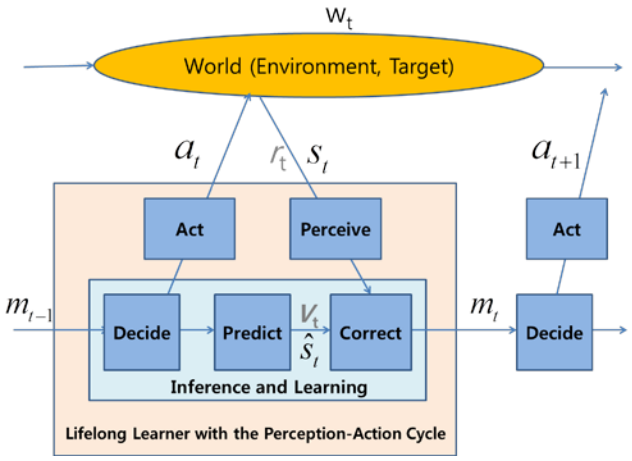

Figure 1: A lifelong learning system architecture with the perception-action cycle

start by considering the information flow in the perception-action cycle of an agent interacting with the environment.

2.1 Action-Perception-Learning Cycle

Consider an agent situated in an environment (Figure 1). The agent has a memory to model the lifelong history. We denote the memory state at time t by m_t. The agent observes the environment and measures the sensory state s_t of the environment and chooses an action a_t. The goal of the learner is to learn about the environment and predict the next world states s_{t+1} as accurately as possible. The ability to predict improves the performance of learner across a large variety of specific behaviors, and is hence quite fundamental, increasing the success rate of many tasks (Still, 2009). The perception-action cycle of the learner is effective for continuous acquisition and refinement of knowledge in a domain or across domains. This paradigm can also be used for time series prediction (Barber et al., 2011), target tracking, and robot motions (Yi et al., 2012). We shall see objective functions for these problems in Sections 3 and 4.

In a different problem setting, the agent is more task-oriented. It has a specific goal, such as reaching a target location or winning a game, and takes actions to achieve the goal. For the actions a_t taken at state s_t, the agent receives rewards r_t from the environment. In this case, the objective is typically formulated to maximize the expected reward $V(s_t)$. The Markov decision problems (Sutton & Barto, 1998) are a representative example of this class of tasks. We shall see variants of objective functions for solving these problems in Section 5.

2.2 Lifelong Learning as Sequential Bayesian Inference

In lifelong learning, the agent starts with the initial knowledge base and continually updates it as it collects more data by observing and interacting with the problem domain or task. This inductive process of evidence-driven refinement of prior knowledge into posterior knowledge can be naturally formulated as a Bayesian inference (Zhang et al., 2012).

The prior distribution of the memory state at time t, is given as $P(m_t^-)$, where the minus sign in m_t^- denotes the memory state before observing the data. The agent collects experience by acting on the environment by a_t and sensing its world state s_t. This action and perception provides the data for computing likelihood $P(s_t, a_t \mid m_t^-)$ of the current model to get the posterior distribution of the memory state $P(m_t \mid s_t, a_t)$. Formally, the update process can be written as

$$
\begin{aligned}
& P(m_t \mid s_t, a_t, m_t^-) \\
&= \frac{P(m_t, s_t, a_t, m_t^-)}{P(s_t, a_t, m_t^-)} \\
&= \frac{1}{P(s_t, a_t, m_t^-)} P(m_t \mid s_t, a_t) P(s_t, a_t \mid m_t^-) P(m_t^-) \\
&\propto P(m_t \mid s_t, a_t) P(s_t, a_t \mid m_t^-) P(m_t^-) \\
&= P(m_t \mid s_t) P(s_t \mid a_t) P(a_t \mid m_t^-) P(m_t^-)
\end{aligned}
$$

where we have used the conditional independence between action and perception given the memory state.

From the statistical computing point of view, a sequential estimation of the memory states would be more efficient. To this end, we formulate the lifelong learning problem as a filtering problem, i.e. estimating the distribution $P(m_t \mid s_{1:t})$ of memory states m_t from the lifelong observations $s_{1:t} = s_1 s_2 ... s_t$ up to time t. That is, given the filtering distribution $P(m_{t-1} \mid s_{1:t-1})$ at time *t-1*, the goal is to recursively estimate the filtering distribution $P(m_t \mid s_{1:t})$ of time step t:

$$
\begin{aligned}
P(m_t \mid s_{1:t}) &= \frac{P(m_t, s_{1:t})}{P(s_{1:t})} \approx P(m_t, s_t, s_{1:t-1}) \\
P(m_t \mid s_{1:t}) &\approx P(s_t \mid m_t) P(m_t \mid s_{1:t-1}) \\
&= P(s_t \mid m_t) \sum_{m_{t-1}} P(m_t, m_{t-1} \mid s_{1:t-1}) \\
&= P(s_t \mid m_t) \sum_{m_{t-1}} P(m_t \mid m_{t-1}) P(m_{t-1} \mid s_{1:t-1})
\end{aligned}
$$

If we let $\alpha(m_t) = P(m_t \mid s_{1:t})$ we have now a recursive update equation:

$$\alpha(m_t) = P(s_t \mid m_t) \sum_{m_{t-1}} P(m_t \mid m_{t-1}) \alpha(m_{t-1})$$

Taking into account the actions explicitly, the recursive lifelong learning becomes:

$$\begin{aligned}
&\alpha(m_t) \\
&= P(s_t \mid m_t) \sum_{m_{t-1}} P(m_t \mid m_{t-1}) \alpha(m_{t-1}) \\
&= \sum_{a_t} P(s_t, a_t \mid m_t) \sum_{m_{t-1}} P(m_t \mid m_{t-1}) \alpha(m_{t-1}) \\
&= \sum_{a_t} P(s_t \mid a_t) P(a_t \mid m_t) \sum_{m_{t-1}} P(m_t \mid m_{t-1}) \alpha(m_{t-1})
\end{aligned}$$

We note that the factors $P(s_t \mid a_t), P(a_t \mid m_t), P(m_t \mid m_{t-1})$ correspond respectively to the perception, action, and the prediction steps in Figure 1. These distributions determine how the agent interacts with the environment to model it and attain novel information.

2.3 Lifelong Supervised Learning

The perception-action cycle formulation above emphasizes the sequential nature of lifelong learning. However, the nonsequential learning tasks, such as classification and regression, can also be incorporated in this framework as special cases. This is especially true for concept learning in a dynamic environment (Zhang et al., 2012). In lifelong learning of classification, the examples are observed as (x_t, y_t), $t = 1, 2, 3, \ldots$, but the examples are independent. The goal is to approximate $\hat{y}_t = f(x_t; m_t)$ with a minimum loss $L(y_q, \hat{y}_q)$ for an arbitrary query input x_q. Note that by substituting

$$\begin{aligned}
s_t &:= x_t \\
a_{t+1} &:= \hat{y}_t = f(x_t; m_t)
\end{aligned}$$

Likewise, the lifelong learning of regression problems can be solved within this framework. The only difference from the classification problem is that in regression the output y_t are real values instead of categorical or discrete values.

3. Learning with Observations

3.1 Dynamical Systems and Markov Models

Dynamical systems involve sequential prediction (Figure 2). For example, time series data consists of $s_{1:T} \equiv s_1, \ldots, s_T$. Since the time length T can be very large or infinite, this problem is an example of lifelong learning problems. In addition, the learner can observe many or indefinite series of different times series, in which case each time series is called an episode.

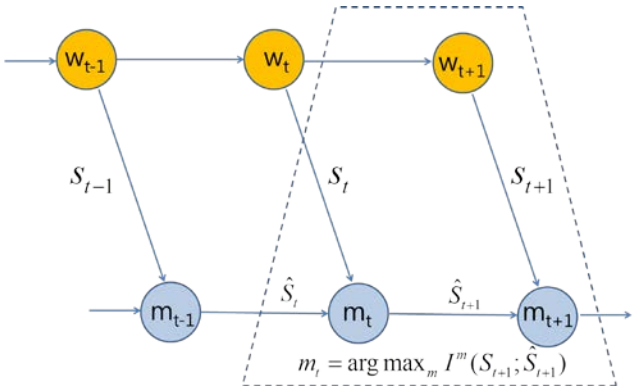

Figure 2: Learning with observations

Dynamical systems can be represented by Markov models or Markov chains. The joint distribution of the sequence of observations can be expressed as

$$P(s_{1:T}) = \prod_{t=1}^{T} P(s_t \mid s_{1:t-1}) = \prod_{t=1}^{T} P(s_t \mid s_{t-1})$$

where in the second equality we used the Markov assumption, i.e. the current state is dependent only on the one previous step.

In time series prediction, the learner has no control over the observations, it just passively receives the incoming data. The goal of the learner is to find the model m_t that best predicts the next state $s_{t+1} = f(s_t; m_t)$ given the current state s_t. How do we define the word "best" quantitatively? In the following subsections we examine three measures: prediction error, predictive information, and information bottleneck. The last two criteria are based on information-theoretic measures.

3.2 Prediction Error

The accuracy of time series prediction can be measured by prediction error, i.e. the mean squared error (MSE) between the predicted states $\hat{s}_{t+1}$ and the observed states s_{t+1}:

$$MSE(s_{1:T}) = \frac{1}{T-1} \sum_{t=1}^{T-1} (s_{t+1} - \hat{s}_{t+1})^2$$

where the prediction $\hat{s}_{t+1}$ is made by using the learner's current model, i.e. $\hat{s}_{t+1} = f(s_t; m_t)$) and n is the length of the series. Thus, a natural measure is the root of the MSE or RMSE and the learner aims to minimize it:

$$\min_m RMSE(s_{1:T}) = \min_m \sqrt{MSE(s_{1:T})}$$

where $m \in M$ is the model parameters.

3.3 Predictive Information

For the evaluation of a time series with an indefinite length, predictive information (Bialek et al., 2001) has been proposed. It is defined as the mutual information (MI) between the future and the past, relative to some instant of t:

$$I(S_{future}; S_{past}) = \left\langle \log_2 \frac{P(s_{future}, s_{past})}{P(s_{future})P(s_{past})} \right\rangle$$

where $\langle\rangle$ symbol denotes an expectation operator. If S is a Markov chain, the predictive information (PI) is given by the MI between two successive time steps.

$$I(S_{t+1}; S_t) = \left\langle \log_2 \frac{P(s_{t+1}, s_t)}{P(s_t)} \right\rangle$$

Several authors have studied this measure for self-organized learning and adaptive behavior. Zahedi et al. (2010), for example, found the principle of maximizing the predictive information effective to evolve a coordinated behavior of the physically connected robots starting with no knowledge of itself or the world. Friston (2009) argues that self-organizing biological agents resist a tendency to disorder and therefore minimize the entropy of their sensory states. He proposes that the brain uses the free-energy principle for action, perception, and learning.

3.4 Information Bottleneck

The information bottleneck method is a technique to compress an unknown random variable X, when a joint probability distribution between X and an observed relevant variable Y is given (Tishby et al., 1999). The compressed variable is Z and the algorithm minimizes the quantity: $\min_{P(z|x)} I(X;Z) - \beta I(Z;Y)$, where $I(X;Z)$ are the mutual information between X and Z.

Creutzig et al. (2009) proposes to use the information bottleneck to find the properties of the past that are relevant and sufficient for predicting the future in dynamical systems. Adapted in our notation, this past-future information bottleneck is written as:

$$\min_{P(m_t|s_t)} \{ I(S_t; \hat{S}_t) - \beta I(\hat{S}_t; S_{t+1}) \}$$

where $S_t, S_{t+1}, \hat{S}_t$ are respectively the input past, the output future, and the model future. Given past signal values a compressed version of the past is to be formed such that information about the future is preserved. When varying β, we obtain the optimal trade-off curve, also known as the information curve, between compression and prediction, which is a more complete characterization of the complexity of the process.

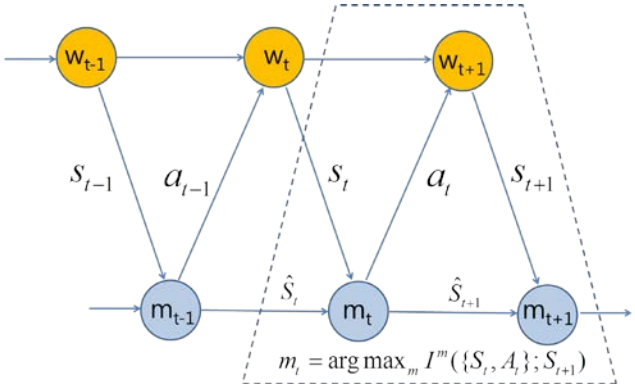

Figure 3: Learning with actions

Creutzig et al. (2009) shows that the past-future information bottleneck method can make the underlying predictive structure of the process explicit, and capture it by the states of a dynamical system. From the lifelong learning point of view, this means that from repeated observations of the dynamic environment the measure provide an objective function that the learner can use to identify the regularity and extract the underlying structures.

4. Learning with Actions

4.1 Interactive Learning

We now consider the learning agents that perform actions on the environment to change the states of the environment (Figure 3). An example of this paradigm is the interactive learning (Still, 2009). Assume that the learner interacts with the environment between consecutive observations. Let one decision epoch consists in mapping the current history h, available to the learner at time t, onto an action (sequence) a that starts at time t and takes time Δ to be executed. The problem of interactive learning is to choose a model and an action policy, which are optimal in that they maximize the learner's ability to predict the world, while being minimally complex.

The decision function, or action policy, is given by the conditional probability distribution $P(a_t \mid h_t)$. Let the model summarize historical information via the probability map $P(s_t \mid h_t)$. The learner uses the current state s_t together with knowledge of the action a_t to make probabilistic predictions of future observations, s_{t+1}:

$$P(s_{t+1} \mid s_t, a_t) = \frac{1}{P(s_t, a_t)} \left\langle P(s_{t+1} \mid h, a_t) P(a_t \mid s_t) P(s_t \mid h) \right\rangle_{P(h)}$$

The interactive learning problem is solved by maximizing $I(\{S_t, A_t\}; S_{t+1})$ over $P(s_t \mid h_t)$ and $P(a_t \mid h_t)$, under constraints that select for the simplest possible

model and the most efficient policy, respectively, in terms of smallest complexity measured by the coding rate. Less complex models and policies result in less predictive power. This trade-off can be implemented using Lagrange multipliers, λ and μ. Thus, the optimization problem for interactive learning (Still, 2009) is given by

$$\max_{P(s_t|h_t),P(a_t|h_t)} \{I(\{S_t, A_t\}; S_{t+1}) - \lambda I(S_t; H_t) - \mu I(A_t; H_t)\}$$

Note that interactive learning is different from reinforcement learning, which will be discussed in the next section. In contrast to reinforcement learning, the predictive model approach such as interactive learning asks about behavior that is optimal with respect to learning about the environment rather than with respect to fulfilling a specific task. This approach does not require rewards. Conceptually, the predictive approach could be thought of as "rewarding" information gain and, hence, curiosity. In that sense, it is related to curiosity driven reinforcement learning (Schmidhuber, 1991, Still & Precup, 2012), where internal rewards are given that correlate with some measure of prediction error. However, the learner's goal is not to predict future rewards, but rather to behave such that the time series that it observes as a consequence of its own actions is rich in causal structure. This, in turn, allows the learner to construct a maximally predictive model of its environment.

4.2 Empowerment

Empowerment measures how much influence an agent has on its environment. It is an information-theoretic generalization of joint controllability (influence on environment) and observability (measurement by sensors) of the environment by the agent, both controllability and observability being usually defined in control theory as the dimensionality of the control/observation spaces (Jung et al., 2012).

Formally, empowerment is defined as the Shannon channel capacity between A_t, the choice of an action sequence, and S_{t+1}, the resulting successor state:

$$\begin{aligned} C(s_t) &= \max_{P(a)} I(S_{t+1}, A_t \mid s_t) \\ &= \max_{P(a)} \{H(S_{t+1} \mid s_t) - H(S_{t+1} \mid A_t, s_t)\} \end{aligned}$$

The maximization of the mutual information is with respect to all possible distribution over A_t. The empowerment measures to what extent an agent can influence the environment by its actions. It is zero if, regardless what the agent does, the outcome will be the same. And it is maximal if every action will have a distinct outcome.

It should be noted that empowerment is fully specified by the dynamics of the agent-environment coupling (i.e. the transition probabilities) and a reward does not need to be specified. Empowerment provides a natural utility function which imbues its states with an a priori value, without an explicit specification of a reward. This enables the system to keep alive indefinitely.

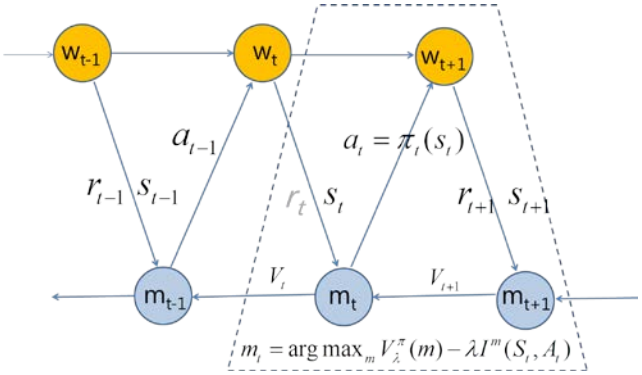

Figure 4: Learning with rewards

5. Learning with Rewards

5.1 Markov Decision Processes

In some settings of lifelong learning, the agent receives feedback information from the environment. In this case, the agent's decision process can be modeled as a Markov decision process (MDP). MDPs are a popular approach for modeling sequences of decisions taken by an agent in the face of delayed accumulation of rewards. The structure of the rewards defines the tasks the agent is supposed to achieve.

A standard approach to solving the MDP is reinforcement learning (Sutton & Barto, 1998), which is an approximate dynamic programming method. The learner observes the states s_t of the environment, take actions a_t on the environment, and gets rewards r_t from it (Figure 4). This occurs sequentially, i.e. the learner observes the next states only after it takes actions. An example of this kind of learner is a mobile robot that sequentially measures current location, takes motions, and reduces the distance to the destination. Another example is a stock-investment agent that observes the state of the stock market, makes sell/buy decisions, and gets payoffs. It is not difficult to imagine extending this idea to develop a lifelong learning agent that incorporates external guidance and feedback from humans or other agents to accumulate knowledge from experience.

5.2 Value Functions

The goal of reinforcement learning is to maximize the expected value for the cumulated reward. The reward

function is defined as $R(s_{t+1} \mid s_t, a_t)$ or $r_{t+1} = r(s_t, a_t)$. This value is obtained by averaging over the transition probabilities $T(s_{t+1} \mid s_t, a_t)$ and the policy $\pi(a_t \mid s_t)$ or $a_t = \pi(s_t)$ Given a starting state s and a policy π, the value $V^\pi(s_t)$ of the state s_t following policy π can be expressed via the recursive Bellman equation (Sutton & Barto, 1998),

$$V^\pi(s_t) = \sum_{a_t \in A} \pi(a_t \mid s_t) \sum_{s_{t+1} \in S} T(s_{t+1} \mid s_t, a_t) \left[R(s_{t+1} \mid s_t, a_t) + V^\pi(s_{t+1}) \right]$$

Alternatively, the value function can be defined on state-action pairs:

$$Q^\pi(s_t, a_t) = \sum_{s_{t+1} \in S} T(s_{t+1} \mid s_t, a_t) \left[R(s_{t+1} \mid s_t, a_t) + V^\pi(s_{t+1}) \right]$$

which is the utility function attained if, in state s_t, the agent carries out action a_t, and after that begins to follow π.

5.3 Information Costs

If there are multiple optimal policies, then asking for the information-theoretically cheapest one among these optimal policies becomes more interesting. Tishby & Polani (2010) and Polani (2011) propose to introduce information cost term in policy learning. It is even more interesting if we do not require the solution be perfectly optimal. Thus, if we wish the expected reward **E**[$V(S)$] to be sufficiently large, the information cost for such as suboptimal (but informationally parsimonious) policy will be generally lower.

For a given utility level, we can use the Lagrangian formalism to formulate the unconstrained minimization problem

$$\min_\pi \left\{ I^\pi(S_t; A_t) - \beta E[Q^\pi(S_t, A_t)] \right\}$$

where $I^\pi(S_t; A_t)$ measures the decision cost incurred by the agent:

$$I^\pi(S_t; A_t) = \sum_{s_t} P(s_t) \sum_{a_t} \pi(a_t \mid s_t) \log \frac{\pi(a_t \mid s_t)}{P(a_t)}$$

where $P(a_t) = \sum_{s_{t+1}} \pi(a_t \mid s_{t+1}) P(s_{t+1})$. The term $I^\pi(S_t; A_t)$ denotes the information that the action A_t carries about the state S_t under policy π.

5.4 Interestingness and Curiosity

The objective function consisting of the value function and the information cost can balance the expected return with minimum cost. However, this lacks any notion of interestingness (Zhang, 1994) or curiosity (Schmidhuber, 1991). In Section 4 we have seen this aspect being reflected in the predictive power and empowerment (Jung et al., 2011). The objective function can be extended by the predictive power (Still & Precup, 2012). Using Lagrange multipliers, we can formulate the lifelong learning as an optimization problem:

$$\arg\max_q \left\{ I_q^\pi \left(\{S_t, A_t\}; S_{t+1} \right) + \alpha V_t^\pi(q) - \lambda I(S_t; A_t) \right\}$$

where $q(a_t \mid s_t)$ is the action policy to be approximated. The ability to predict improves the performance of a learner across a large variety of specific behaviors.

The above objective function embodying the curiosity terms as well as the value and information cost terms can thus be an ideal guideline for a lifelong learner. The predictive power term $I_q^\pi \left(\{S_t, A_t\}; S_{t+1} \right)$ allows for the agent to actively explore the environment to extract interesting knowledge. The information cost term $I(S_t; A_t)$ enables the learner to minimize the interaction with the environment or teacher. This all happens with the goal of maximizing the value or utility $V_t^\pi(q)$ of the information the agent is acquiring.

6. Summary and Conclusion

We have formulated lifelong learning as a sequential, online, incremental learning process over an extended period of time in a dynamic, changing environment. The hallmark of this lifelong-learning framework is that the training data are observed sequentially and not kept for iterative reuse. This requires instant, online model building and incremental transfer of knowledge acquired from previous learning to future learning, which can be formulated as a Bayesian inference.

The Bayesian framework is general enough to cover the perception-action cycle model of cognitive systems in its various instantiations. We applied the framework to develop a taxonomy of lifelong learning based on the way of obtaining learning examples. We distinguished three paradigms: learning with observations, learning with actions, and learning with rewards. For each of the paradigms we examined the objective functions of the lifelong learning styles.

The first paradigm is lifelong learning with passive, continual observations. Typical examples are time series prediction and target tracking (filtering). The objective functions for this setting are prediction errors and predictive information, the latter being defined as the

mutual information between the past and future states in time series. The information bottleneck method can also be modified to measure the predictive information.

The second paradigm is lifelong learning with actions (but without reward feedbacks). Interactive learning and empowerment are the examples. Here, the learner actively explores the environment to achieve maximal predictive power at minimal complexity about the environment. In this paradigm, the agent takes actions on the environment by action policy, but does not receive rewards from the environment for its actions on the environment. The goal is mainly to know more about the world. Simultaneous localization and mapping (SLAM) in robotics is an excellent example of the interactive learning problem, though no literature is found on explicit formulation of SLAM as interactive learning.

The third paradigm is active lifelong learning with explicit rewards. This includes the MDP problem for which approximate dynamic programming and reinforcement learning have been extensively studied. The conventional objective function for MDPs is the value function or the expected reward of the agent. As we have reviewed in this paper, there have been several proposals recently to extend the objective function by incorporating information-theoretic factors. These objective functions can be applied to lifelong learning agents, for example, to attempt to minimize information costs while maximizing the predictive information or curiosity for a given level of expected reward from the environment. These approaches are motivated by information-theoretic analysis of the perception-action cycle view of cognitive dynamic systems.

In this article, we have focused on the sequential, predictive learning aspects of lifelong learning. This framework is general and thus can incorporate the classes of lifelong classification and regression learning. Since these supervised learning problems do not care about the sequence of observations, the sequential formulations presented in this paper can be reused by ignoring the temporal dependency. We also did not discuss the detailed mechanisms of learning processes for the lifelong learning framework. Future work should relate the information-theoretic objective functions to the representations to address questions like "how to discover and revise the knowledge structures to represent the internal model of the world or environment" (Zhang, 2008).

As a whole, we believe the general framework and the objective functions for lifelong learning described here provide a baseline for evaluating the representations and strategies of the learning algorithms. Specifically, the objective functions can be used for innovating algorithms for discovery, revision, and transfer of knowledge of the lifelong learners over the extended period of experience. Our emphasis on information theory-based active and predictive learning with minimal mechanistic assumptions on model structures can be especially fruitful for automated knowledge acquisition and sequential knowledge transfer between a wide range of similar but significantly different tasks and domains.

Acknowledgements: This work was supported in part by the National Research Foundation (NRF-2010-0017734) and the AFOSR/AOARD R&D Grant 124087.

References

[Ay et al., 2008] Ay, N., Bertschinger, N., Der, R., Guetter, F., & Olbrich, E., Predictive information and explorative behavior in autonomous robots, *European Physical Journal* B, 63:329-339, 2008.

[Barber et al., 2011] Barber, D., Cemgeil, A. T., & Chiappa, S. (eds.), *Bayesian Time Series Models*, Cambridge University Press, 2011.

[Bialek et al., 2001] Bialek, W., Nemenman, I., & Tishby, N., Predictability, complexity, and learning, *Neural Computation*, 13:2409-2463, 2001.

[Cohn et al., 1990] Cohn, D., Atlas, L., & Ladner, R., Training connectionist networks with queries and selective sampling, In: D. Touretzky (ed.), *Advances in Neural Information Processing 2*, Morgan Kaufmann, 1990.

[Cohn et al., 1994] Cohn, D., Atlas, L., & Ladner, R., Improving generalization with active learning, *Machine Learning*, 15(2):201-221, 1994.

[Creutzig et al., 2009] Creutzig, F., Globerson, A., & Tishby, N., Past-future information bottleneck in dynamical systems, *Physical Review E*, 79, 042519, 2009.

[Eaton & desJardins, 2011] Eaton, E. & desJardins, M., Selective transfer between learning tasks using task-based boosting, In: *Proc. Twenty-Fifth AAAI Conf. Artificial Intelligence (AAAI-11)*, pp. 337–342, AAAI Press, 2011.

[Freund et al., 1993] Freund, Y., Seung, H. S., Shamir, E., & Tishby, N., Information, prediction, and query by committee, In: S. Hanson et al. (eds)., *Advances in Neural Information Processing 5*, Morgan Kaufmann.

[Friston, 2009] Friston, K., The free-energy principle: a unified brain theory?, *Nature Reviews Neuroscience*, 11:127-138, 2009.

[Jung et al., 2011] Jung, T., Polani, D., & Stone, P., Empowerment for continuous agent-environment systems, *Adaptive Behavior*, 19(1):16-39, 2011.

[Polani, 2011] Polani, D., An information perspective on how the embodiment can relieve cognitive burden, In *Proc. IEEE Symposium Series in Computational Intelligence: Artificial Life*, IEEE Press, pp. 78-85, 2011.

[Schmidhuber, 1991] Schmidhuber, J., Curious model-building control systems, In *Proc. Int. Joint. Conf. Neural Networks*, pp. 1458-1463, 1991.

[Still, 2009] Still, S., Information-theoretic approach to interactive learning, *European Physical Journal*, 85, 2009.

[Still & Precup, 2012] Still, S. & Precup, D., An Information-theoretic approach to curiosity-driven reinforcement learning, *Theory in Biosciences*, 131(3):139-148, 2012.

[Sutton & Barto, 1998] Sutton, R. S. & Barto, A. G., Reinforcement Learning: An Introduction, MIT Press, 1998.

[Tishby et al., 1999] Tishby, N., Pereira, F. C., & Bialek, W., The information bottleneck method, In: *Proc. 37th Annual Allerton Conf. Communication, Control and Computing*, 1999.

[Tishby & Polani, 2010] Tishby, N. & Polani, D., Information theory of decisions and actions. In: *Perception-Reason-Action Cycle: Models, Algorithm and Systems*. Springer, 2010.

[Thrun & Moeller, 1992] Thrun, S. & Moeller, K. Active exploration in dynamic environments, In: J. Moody et al., (eds.) *Advances in Neural Information Processing 4*, Morgan Kaufmann, 1992.

[Valiant, 1984] Valiant, L. G., A theory of the learnable, *Communications of the ACM*, 27(11):1134-1342, 1984.

[Yi et al., 2012] Yi, S.-J., Zhang, B.-T., & Lee, D. D., Online learning of uneven terrain for humanoid bipedal walking, In *Proc. AAAI Conference on Artificial Intelligence (AAAI 2010)*, pp. 1639-1644, 2010.

[Zahedi et al., 2010] Zahedi, K., Ay, N., & Der, R., Higher coordination with less control – A result of information maximization in the sensorimotor loop, Adaptive Behavior, 18(3-4):338-355, 2010.

[Zhang et al., 2012] Zhang, B.-T., Ha, J.-W., & Kang, M., Sparse population code models of word learning in concept drift, In: *Proc. 34th Annual Conference of the Cognitive Science Society (CogSci 2012)*, pp. 1221-1226, 2012.

[Zhang, 2008] Zhang, B.-T., Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, 3(3):49-63, 2008.

[Zhang, 1994] Zhang, B.-T., Accelerated learning by active example selection, *International Journal of Neural Systems*, 5(1):67-75, 1994.

[Zhang & Veenker, 1991a] Zhang B.-T. & Veenker, G., Focused incremental learning for improved generalization with reduced training sets, *Proc. Int. Conf. Artificial Neural Networks (ICANN'91)*, pp. 227-232, 1991.

[Zhang & Veenker, 1991b] Zhang B.-T. & Veenker, G., Neural networks that teach themselves through genetic discovery of novel examples, *Proc. 1991 IEEE Int. Joint Conf. Neural Networks (IJCNN'91)*, pp. 690-695, 1991.